Module 5: Data Visualization

Case Study

Things you will learn in this case-study:

1. Install and load “dplyr” as well as ‘ggplot2’ package
2. Understand Grammar of Graphics
3. Play with different components of grammar of graphics
4. Build cool visualization on your data

Back Ground:

Real Estate Companies like Godrej Reality spend billions on customer acquisition and retention. This spending is in the areas of marketing, advertisement, sales staff etc. They are facing a lot of competition in the sector and with all the regulatory compliances coming into place, the task has become even more daunting.
As a result, these companies are using analytics to understand their data about customers and transactions to have a focused approach towards the entire process. This is where they require your support.

Overview of the problem:

In this project, you will play the role of Data Scientist for the Real Estate Company named, Housing Ltd, and you have been asked to create visualization of all the important transactional variables. Post this, the Analytics Manager will get the visual of the key features and consider it in their next planning review.
The Data you will be dealing is with Transactions Data captured for the Company. The data is in data frame format which has been stored in the file named ‘housingdata.csv’

Data set description:

Record: Record ID
Gender: Gender of the applicant.
No_kids: Number of kids.
Education: Education level of the individual.
HasCar: whether the individual has car or not.
Income: income level of the individual.
PropertyValue: Property Value of the Flat / House (in USD)
Loan_Period: period of the loan (in months)
Credit_Record: value of 1, if the Credit Record is good and 0, if it is not.
Housing_type: category of housing property – (Affordable, Mid-Range and Premium)

Property_Purchased: takes the value 'Y', if the inquiry led to Purchase of the property, and 'N' if the property was not Purchased.

Objective:

Based on the knowledge you acquired in Module 5, you are expected to complete the below mentioned activities by using the ggplot2 package, which is the most widely used library in R for data visualization.

You should do the following:

1. Load the required libraries and the data.

2. Understand the data structure and provide concise summary on the following –
   - no of observations
   - total number of variables
   - number of continuous variables
   - number of categorical variables
   - number of variables which have missing values

3. Create a scatter plot between Credit_Record on x-axis and Income on y-axis.
   - Is the plot satisfying, if not, what could be the reason?
   - Change the command executed in the previous line so that Credit_Record is treated as factor.
   - what is the change in the above two plots?

4. Create a scatter plot between Income on x-axis and PropertyValue on y-axis.
   - In the above plot, add the color argument which should be dependent on the No_kids of the applicant
   - In the above plot, now add the size argument which should be dependent on the No_kids of the applicant.
   - Now, in the above plot, please add the smooth line using the geom_smooth() function.

5. ggplot comparson with Base plot :
   - Using the base package plot(), make a scatter plot with Income on the x-axis and PropertyValue on the y-axis, colored according to No of kids (use the col argument).
   - Now, Change No_kids in previous step to a factor
   - Now, Make the same plot as in the first instruction - 5a
   - Now, recreate the same plot as above using the ggplot functon.

6. Aesthetics:
   - Map Income to x and Property Value to y
   - Reverse: Map Property Value to x and Income to y
   - Map Income to x and Property Value to y and No of kids to col
   - Change shape and size of the points in the above plot.

7. Geometry:
   - Start with creating a scatter plot mapping Income to x and Property Value to y.
   - Make a plot With geom_jitter() function
   - Now, in the above plot, Set width in geom_jitter(). Take the width value as 0.1

8. Histogram:
   - Make a univariate histogram on Income
   - In the above plot, add set binwidth to 100 in the geom layer
   - In the above plot, MAP ..density.. to the y aesthetic (i.e. in a second aes() function)
   - Finally, in the above plot, plus SET the fill attribute to "#377EB8".

9. Bar Plot:
   - Draw a bar plot of Property_Purchased, filled according to Education
   - In the above plot, Change the position argument to "stack"
   - In the above plot, Change the position argument to "fill"
   - In the above plot, Change the position argument to "dodge"

10. Overlapping bar plots:
    - Take the last plot form the previous exercise
    - In the above plot, Define posn_d with position_dodge(). Take value as 0.7
    - Change the position argument to posn_d in the last plot made in Step 9(d)
    - Use posn_d as position and adjust alpha to 0.6 - can you see the overlap in bars. If not, change the value of alpha
    -

11. Overlapping histograms:
    - A basic histogram, add coloring defined by Income and filled by HasCar, select a suitable binwidth
    - In the above plot, In the above chart, Change position to identity

12. Faceting:
    - Now create a basic scatter plot between income and property value variables
    - In the above plot, Separate rows according to HasCar

- In plot made in step 12b, Separate columns according to No of kids
- In plot made in step 12b, , Separate by both HasCar and No of kids

Submission should include the following:

1. Answers to the above questions. Print the resultant plots as output wherever applicable.
2. Summary on approach should be documented and submitted for each question.
3. R Code File.